(12)特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2005 年8 月11 日 (11.08.2005)

PCT

(10) 国際公開番号
WO 2005/073874 A1

(51) 国際特許分類[7]: G06F 17/28

(21) 国際出願番号: PCT/JP2005/001636

(22) 国際出願日: 2005 年1 月28 日 (28.01.2005)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願2004-023913 2004 年1 月30 日 (30.01.2004) JP

(71) 出願人 (米国を除く全ての指定国について): 独立行政法人情報通信研究機構 (NATIONAL INSTITUTE OF INFORMATION AND COMMUNICATIONS TECHNOLOGY, INDEPENDENT ADMINISTRATIVE AGENCY) [JP/JP]; 〒1840015 東京都小金井市貫井北町4-2-1 Tokyo (JP).

(72) 発明者; および

(75) 発明者/出願人 (米国についてのみ): 内元 清貴 (UCHIMOTO, Kiyotaka) [JP/JP]; 〒1840015 東京都小金井市貫井北町4-2-1 独立行政法人情報通信研究機構内 Tokyo (JP). 井佐原 均 (ISAHARA, Hitoshi) [JP/JP]; 〒1840015 東京都小金井市貫井北町4-2-1 独立行政法人情報通信研究機構内 Tokyo (JP).

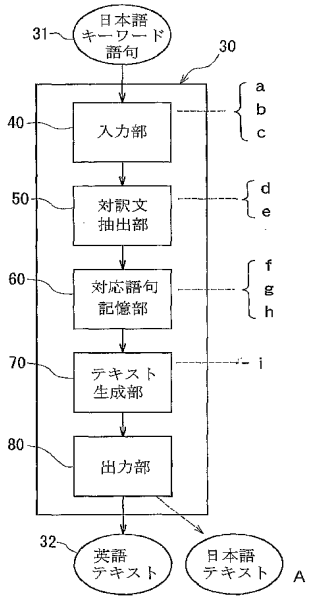
(74) 代理人: 渡邊 敏 (WATANABE, Satoshi); 〒1600008 東京都新宿区三栄町18-20 パークサイド四谷２階 渡辺特許法律事務所内 Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG,

/続葉有/

(54) Title: OTHER LANGUAGE TEXT GENERATION METHOD AND TEXT GENERATION DEVICE

(54) 発明の名称: 他言語のテキスト生成方法及びテキスト生成装置



31 JAPANESE KEYWORD PHRASE
40 INPUT UNIT
50 TRANSLATION SENTENCE EXTRACTION UNIT
60 CORRESPONDING PHRASE STORAGE UNIT
70 TEXT GENERATION UNIT
80 OUTPUT UNIT
32 ENGLISH TEXT
A JAPANESE TEXT

(57) Abstract: By inputting words of an original language as a keyword (31), a translation sentence is extracted (50) from a translation corpus database between an original language and another language. From the partially corresponding information on the translation sentence, a corresponding phrase group table formed by the corresponding phrase of the another language corresponding to the original language phrase including a keyword phrase of the original language is stored (60). Text generation means (70) assumes a relationship between the phrases of different languages contained in the corresponding phrase group table and generates a text candidate (32) of the another language.

(57) 要約: 原言語の単語をキーワード３１として入力することにより、原言語・他言語間の対訳コーパスデータベースから抽出する対訳文抽出５０し、該対訳文の部分対応情報から、各原言語のキーワード語句を含む原言語対応語句に対応する他言語の各他言語対応語句で構成する対応語句群テーブルを記憶６０する。テキスト生成手段７０では、該対応語句群テーブルに含まれる各他言語対応語句間の係り受け関係を仮定して他言語のテキスト候補３２を生成する。

WO 2005/073874 A1

SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:
— 国際調査報告書

２文字コード及び他の略語については、定期発行される各*PCT*ガゼットの巻頭に掲載されている「コードと略語のガイダンスノート」を参照。

明細書

他言語のテキスト生成方法及びテキスト生成装置

技術分野

本発明は自然言語処理の方法及び装置に関する。特に、原言語の単数又は複数のキーワード語句から他言語のテキストを生成する手法に関する。

背景技術

計算機を用いてテキストを解析、生成するための方法は従来から数多く提案されている。それらを大別すると、人間が作成した規則に基づく方法と統計的学習に基づく方法に分けることができる。前者の方法では、多様な知識を利用することで処理精度を向上させようとしてきた。一方、後者の方法では、単純な知識を大量に利用することで処理精度を向上させようとしてきた。

テキストを精度良く解析、生成するためには、文内、文間に現われる表層的情報から得られる様々な知識をはじめとして、辞書的な知識、言語学的な知見など、できるだけ多様な知識を利用するのが良いと考えられる。

しかし、前者の方法では、多様な知識を扱うためには規則を精緻化しなくてはならず、必然的に規則が競合しやすくなり、規則同士の優先順位を決めるのが困難になる。

一方、後者の方法では、多様な知識を利用しようとすると学習データに過学習する傾向があるため、過学習を避けるためにさらに多くの学習データが必要となることが多い。後者の方法で多様な知識を利用することができればより良い精度が期待できる。しかし、後者の方法では、これまで知識を充実させるという方向の研究はほとんどなされてこなかった。

本件発明者らは、後者の統計的学習に基づく方法を採用し、テキスト解析・生成のための新しいモデルを提案しており、例えば、特開２００２－３３４０７６号公報において開示している。このモデルは、主に最大エン

トロピー原理に基づくもので、過学習の問題を避けつつ、多様な知識を効率良く扱うことができる。実験により、既存の統計的方法に比べて高い精度が得られることを示すとともに、学習データから得られる知識や、辞書的な知識、言語学的な知見などの多様な知識を効率的に利用する方法、および、テキスト解析・生成に有効な知識とはどのようなものであるかが明らかになっている。

一方、具体的なテキスト生成の処理方法としては、例えば本件出願人による特開２００３－１９６２８０号公報に開示されるテキスト生成のシステムがある。該システムでは、キーワードを入力してそれを含むテキストをデータベースから抽出し、該テキストを形態素解析・構文解析した後、もとのキーワードをテキストに組み合わせることでテキストの生成を行うように処理している。

また、特開２００３－２７１５９２号公報に開示されたシステムでは、キーワードとなる単語を入力して、文字単位候補を生成し、文字単位候補の係り受け関係を仮定してテキスト候補を生成するテキスト生成方法を開示している。本方法によると、キーワードが十分でない場合にも自然なテキストを生成できる長所がある。

これらはいずれも、例えば日本語のキーワードから日本語のテキストを生成するものであって、異なる言語のテキストを生成する手法ではない。すなわち従来の技術では単言語のコーパスを用いて、単言語のキーワードからテキスト生成する方法が提供されているだけであり、上記特開２００３－２７１５９２号公報の方法を他言語に適用する方法は実現できていなかった。

また、入力する言語と出力する言語が異なる言語処理としては機械翻訳が知られている。機械翻訳の一般的な手法は、翻訳元言語のテキストを入力し、それを解析、その解析結果から翻訳先言語を生成する。

しかし、入力時に必ずしもテキストを入力せず、適当なキーワードを与えることで、より自然なテキストを出力できるのであれば、使用者にとって他者とのコミュニケーションをより図りやすくなることも考えられる。

例えば近年、ネットワークを通じて世界中の人々が容易に情報を交換できるようになったが、依然として言語バリアが存在しており、異文化間のコミュニケーションは容易ではない。これまでに、機械翻訳の技術は向上してきたが、商用の機械翻訳システムを用いてもなお異文化間のコミュニケーションは難しいということが指摘されている。

そこで、異文化間コミュニケーションにおける言語バリアを克服するために、システムに対する人間の協調をうまく引き出し、異文化間コミュニケーションを可能とするような他言語のテキスト生成方法が求められている。

本発明は、上記従来技術の有する問題点に鑑みて創出されたものであり、その目的は、使用者が適当なキーワード語句を与えることによりごく自然な他言語のテキスト生成を実現する他言語テキスト生成方法及び他言語テキスト生成装置を提供することである。

発明の開示

本発明は、上記の課題を解決するために、次のような他言語テキスト生成方法を創出する。

すなわち、原言語の語句をキーワード語句として入力することにより、原言語とは異なる他言語のテキストを生成する他言語テキスト生成方法である。そして、入力ステップにおいて、入力手段から、単数又は複数の該原言語のキーワード語句を入力する。対訳文中の語句間対訳関係に係る部分対応情報を含む原言語・他言語間の対訳コーパスデータベースを用い、対訳文抽出手段が、入力されたキーワード語句を含む対訳文を、該対訳コーパスデータベースから抽出する対訳文抽出ステップ、該対訳文の部分対応情報から、各原言語のキーワード語句を含む原言語対応語句に対応する他言語の各他言語対応語句で構成する対応語句群テーブルを記憶手段に記憶する対応語句記憶ステップ、テキスト候補生成手段が、該対応語句群テーブルに含まれる各他言語対応語句間の係り受け関係を仮定して他言語のテキスト候補を生成するテキスト候補生成ステップを備える。最後に出力

ステップで、出力手段から、少なくとも１つのテキスト候補を出力する。

ここで、上記他言語テキスト生成方法の対訳文抽出ステップにおいて、入力ステップで入力したキーワード語句に対して、複数の対訳文が抽出され、部分対応情報から原言語対応語句が複数の種類存在するときに、対訳文抽出ステップの次に、複数の原言語対応語句を使用者に選択可能に提示する原言語語句候補提示ステップを備え、対応語句記憶ステップにおいて、使用者が選択した場合に、その原言語対応語句に対応する他言語対応語句を対応語句群記憶テーブルに記憶するようにしてもよい。

また、入力ステップにおいて、１個のキーワード語句を入力する毎に、対訳文抽出ステップ及び、前記対応語句記憶ステップの各処理を行うと共に、抽出された対訳文中において該キーワード語句と共起する共起語句を抽出し共起語句テーブルに記憶する共起語句抽出ステップ、該共起語句テーブル中の共起語句を使用者に選択可能に提示する共起語句提示ステップの各ステップを備える。入力ステップにおいて、使用者が共起語句を選択した場合には、その共起語句を新たなキーワード語句として入力し、全てのキーワード語句の入力が終了した後に、テキスト候補生成ステップに進むように構成してもよい。

さらに、他言語テキスト生成方法の対訳文抽出ステップにおいて、各処理に先だち、入力ステップで入力されたキーワード語句について、構成する一部の形態素の加減、又は類語への置換を行うようにしてもよい。

また、他言語テキストが複数の言語であって、対訳文抽出ステップ、対応語句記憶ステップ、テキスト候補生成ステップにおいて、原言語と、各他言語との間についてそれぞれ処理を行うものである。例えば、他言語テキストとして、第１言語、第２言語、第３言語がある場合には、原言語と第１言語、原言語と第２言語、原言語と第３言語の間でそれぞれ上記処理を行う。これにより、出力ステップにおいては、全ての他言語のテキスト候補を出力するように構成してもよい。

さらに、テキスト候補生成ステップにおいて、テキスト候補生成手段が、該対応語句群テーブルに含まれる各他言語対応語句間の係り受け関係を

仮定して他言語のテキスト候補を生成すると共に、原言語テキスト候補生成手段が、該対応語句群テーブルに含まれる各原言語対応語句間の係り受け関係を仮定して原言語のテキスト候補を生成する。そして、出力ステップにおいて出力手段から、少なくとも１組の原言語及び他言語の対訳テキスト候補を共に出力するようにしてもよい。

また、テキスト候補生成ステップの次に、評価手段が、該テキスト候補を評価付けする評価ステップを有し、出力ステップにおいては、該評価に基づいて少なくとも１つのテキスト候補を出力するように構成してもよい。

また、本発明は、上記課題を解決するために、次のような他言語テキスト生成装置を創出する。

すなわち、原言語の単語をキーワードとして入力することにより、原言語とは異なる他言語のテキストを生成する他言語テキスト生成装置であって、単数又は複数の該原言語のキーワード語句を入力する入力手段と、対訳文中の語句間対訳関係に係る部分対応情報を含む原言語・他言語間の対訳コーパスデータベースと、該キーワード語句を含む対訳文を、該対訳コーパスデータベースから抽出する対訳文抽出手段と、該対訳文の部分対応情報から、各原言語のキーワード語句を含む原言語対応語句に対応する他言語の各他言語対応語句で構成する対応語句群テーブルを記憶可能な対応語句記憶手段と、該対応語句群テーブルに含まれる各他言語対応語句間の係り受け関係を仮定して他言語のテキスト候補を生成するテキスト候補生成手段と、少なくとも１つのテキスト候補を出力する出力手段とを少なくとも備える。

ここで、入力したキーワード語句に対して前記対訳文抽出手段により複数の対訳文が抽出され、その部分対応情報から原言語対応語句が複数の種類存在するか否か判定し、複数の種類存在する場合には、使用者に該各原言語対応語句を提示する原言語語句候補提示手段を備えると共に、前記入力手段から、使用者が提示された原言語対応語句の１個を選択可能であり、使用者が選択した場合には、前記対応語句記憶手段がその原言語対応語

句に対応する他言語対応語句を対応語句群記憶テーブルに記憶するように構成してもよい。

また、入力手段から１個のキーワード語句を入力する毎に、前記対訳文抽出手段及び、前記対応語句記憶手段が作用する構成において、抽出された対訳文中において該キーワード語句と共起する共起語句を抽出し共起語句テーブルに記憶する共起語句抽出手段と、該共起語句テーブル中の共起語句を使用者に選択可能に提示する共起語句提示手段とを備える。そして、入力手段から使用者が共起語句を選択した場合には、該共起語句を新たなキーワード語句として入力し、全てのキーワード語句の入力が終了した後に、前記テキスト候補生成手段が作用するようにしてもよい。

さらに、入力手段から入力されたキーワード語句について、構成する一部の形態素の加減、又は類語への置換を行うキーワード整形手段を備え、対訳文抽出手段において処理を行うようにしてもよい。

また、対訳コーパスデータベースに、原言語と、複数の他言語との間の対訳文中の語句間対訳関係に係る部分対応情報を含み、対訳文抽出手段と、対応語句記憶手段と、テキスト候補生成手段において、該原言語と、各他言語との間についてそれぞれ処理を行うと共に、出力手段から、複数の言語のテキスト候補を出力するようにしてもよい。

さらに、テキスト候補生成手段が、前記対応語句群テーブルに含まれる各他言語対応語句間の係り受け関係を仮定して他言語のテキスト候補を生成すると共に、該対応語句群テーブルに含まれる各原言語対応語句間の係り受け関係を仮定して原言語のテキスト候補を生成する原言語テキスト候補生成手段を備え、出力手段から、少なくとも１組の原言語及び他言語の対訳テキスト候補を共に出力するようにしてもよい。

また、テキスト候補を評価付けする評価手段を備えるようにしてもよい。

図面の簡単な説明

第１図は、本発明で用いるコーパスの依存構造木の説明図である。

第２図は、本発明の実施例１に係るテキスト生成方法のフローチャートである。

第３図は、本発明の実施例２に係るテキスト生成方法のフローチャートである。

第４図は、本発明の実施例３に係るテキスト生成方法のフローチャートである。

第５図は、本発明の実施例４に係るテキスト生成方法のフローチャートである。

第６図は、本発明の実施例５に係るテキスト生成方法のフローチャートである。

第７図は、本発明のテキスト生成装置の構成図である。

第８図は、本発明における入力部の構成図である。

第９図は、本発明における対訳文抽出・対応語句記憶部の構成図である。

第１０図は、本発明におけるテキスト生成部の構成図である。

第１１図は、英語対応語句からのテキスト生成の例を示す説明図である。

第１２図は、英語対応語句と単語列との関係を示す説明図である。

第１３図は、本発明におけるテキスト生成部（実施例３）の構成図である。

第１４図は、本発明の実施例６に係るテキスト生成方法のフローチャートである。

第１５図は、本発明におけるテキスト生成部（実施例６）の構成図である。

発明を実施するための最良の形態

以下、本発明を実施するための最良の形態を、図面に示す実施例を基に説明する。なお、本発明はこの実施の形態に限定されるものではない。

まず、本発明の要部につき説述する。従来から母国語などを入力して異

なる言語のテキスト（文章又はその集合）を出力する機械翻訳技術は知られており、近年高精度な機械翻訳が可能になりつつある。しかしながら、原言語のテキストを解析する過程と、他言語のテキストを生成する過程それぞれで、それぞれの言語が有する自然な言い回しや語順などが崩れてしまう場合があり、翻訳としては誤りではなくとも、コミュニケーションを図るために最適なテキストを得ることは難しい問題があった。

また、機械翻訳の性能が十分に高くないと、原言語の入力時に機械翻訳に適する言い回しに直して入力しなければならなかったり、必要な言葉を過不足無く入力文に盛り込まなければならなかったりして、誰にでも簡便に使用することは難しい。一方で、インターネットの普及により世界中の誰とでも気軽にコミュニケーションをとれるようになった昨今において、正しいニュアンスの他言語を生成し、コミュニケーションを図れるような支援方法の提供は急務である。

そこで本発明では母国語などのキーワード語句をいくつか入力することで、該キーワード語句の対訳語句を用いる他言語のテキストを生成する方法を創出した。使用者は母国語で伝えたい内容のうち重要な単語等を入力することにより、装置があらかじめ備えている対訳テキストのデータベースからそれらを用いる他言語テキストが生成される。その上、伝えている内容は原言語で確認できるため、使用者は正確なニュアンスの他言語テキストが生成されているかを確認することができる。

この方法で用いる対訳コーパスと呼ばれるデータベースは、原言語と他言語のそれぞれの文が対訳関係を持って格納されており、最初は人手によって正確な翻訳文を作成することが望ましい。そして、それぞれの文には構文情報も付与されており、句のレベルでの言語間の対応も付与されている。

本件発明者らが開発している対訳コーパスの１つとして、日本語と英語の対訳コーパスが完成しており、該コーパスは新聞記事を基にプロの翻訳家により作成したもので、日英文数は現在約４万である。

本コーパスは、英訳は日本文１文に対して１つの訳文（1 文）とし、自

然な英文に訳出してある。日本文で主語が省略されている場合は、前文章の流れで必要に応じて主語を補い、主語に代名詞を持ってくるか、固有名詞かは前文からの自然な流れで決定する。このように作出するため、本コーパスは日本文・英文共に自然な言葉で表現されている。

コーパスのデータ形式を簡単に説述する。例えば日本文で、「また、一九九五年中の衆院解散・総選挙の可能性に否定的な見解を表明、二十日招集予定の通常国会前の内閣改造を明確に否定した。」に対して、図１のような依存構造木を定義し、依存構造木の左側に文節毎に付したＩＤを用いて、

＊ 0 12D
また また ＊ 接続詞 ＊ ＊
、 、 ＊ 特殊 読点 ＊ ＊
＊ 1 2D
一九九五 いちきゅうきゅうご ＊ 名詞 数詞 ＊ ＊
年 ねん ＊ 接尾辞 名詞性名詞助数辞 ＊ ＊
中 ちゅう ＊ 接尾辞 名詞性名詞接尾辞 ＊ ＊
の の ＊ 助詞 接続助詞 ＊ ＊

というように順に文節の番号、係り受け先、形態素、読み、品詞などを定義する。

さらに、この対訳文「He also responded negatively to the possibility of dissolution of the House of Representatives and general elections before the end of 1995, and clearly denied a cabinet reshuffle would take place prior to the ordinary Diet session scheduled to be convened on the 20th.」について、「<P id="6,7">He<¥P> <P id="1">also<¥P> <P id="6,7">responded<¥P> <P id="5">negatively<¥P> <P id="4">to the possibility<¥P> <P id="3">of dissolution of the House of Representatives and general elections<¥P>・・・」

と、上記日本文の文節ＩＤをタグ（<P id="">と<¥P>で囲まれた部分）で表示しながら、各ワードの部分対応情報としている。

（実施例１）

図２には本発明による実施例１に係る他言語テキスト生成方法のフローチャートを示す。図のように、原言語（日本語）のキーワード語句（１）を入力し、そのキーワード語句（１）を含む対訳文を、対訳コーパスデータベース（１０）から抽出（２）する。

そして、対訳文中からキーワード語句に関係する対応語句を部分対応情報（１１）から抽出し、対応語句群テーブル（１２）として記憶する。なお、該部分対応情報（１１）は実際には対訳コーパスデータベース（１０）中に含まれている情報であるから、両データは一体である。

ここまでの処理によって入力したキーワード語句に対応する他言語の語句が得られる。この後、これらの語句間の係り受け関係を仮定して他言語のテキスト候補を生成（４）する。

得られたテキスト候補はそのまま出力する構成でもよいが、本実施例ではこの後これらを評価（５）し、候補の中から最も適当な他言語（英語）テキスト（６）を出力する。

次に、本発明による他言語テキスト生成方法を実現するための他言語テキスト生成装置の構成を図７に示す。本装置（３０）は、例えば「彼女」「公園」「行く」などの日本語キーワード語句を入力すると、入力部（４０）で装置（３０）内への取り込み処理を行い、対訳文抽出部（５０）において「公園へ行った／I went to the park」「彼女と百貨店へ行った／I went to the department store with her」などの対訳文が抽出される。

さらに対応語句記憶部（６０）で、部分対応情報から上記対訳文の中でキーワード語句に関する「公園へ／to the park」「行った／I went ...」「彼女と／with her」などが抽出され、記憶する。

テキスト生成部（７０）では、これらの対応語句から「I went to the park with her」という英語のテキストを生成し、出力部（８０）から英語テキスト（３２）を出力する。

次に各部（４０）ないし（８０）の詳細を説述する。入力部（４０）は図８に示すようにＣＰＵ（４１）とそれに接続されたマウス（４２）やキ

ーボード（４３）、ＣＤドライブ、ハードディスクドライブ、ＭＯドライブ、フロッピー（登録商標）ディスクドライブなどの記憶装置（４４）等から構成される。また、ＣＰＵ（４１）の動作に伴い、必要に応じて公知のメモリを用いることもできる。

使用者はマウス（４２）やキーボード（４３）により直接キーワード語句を入力することができる。

また、本発明はインターネットやイントラネットのネットワーク（４５）を介して他のコンピュータサーバー等からキーワード語句を受信することも可能である。

公知のタッチパネルモニタ（４６）を設けてより簡便な入力方法を提供してもよい。

入力部（４０）により日本語キーワード語句（３１）は図９に示される対訳文抽出（５０）・対応語句記憶（６０）部に送られる。

本実施例では、対訳文抽出（５０）・対応語句記憶（６０）部は１個の処理部（５１）として図示する。ここでもＣＰＵ及びメモリが協働して各処理を行う。

まず対訳文抽出部（５０）は外部記憶装置に格納された対訳コーパスデータベース（５２）から日本語キーワード語句（３１）を文中に含む対訳文を抽出する。

このとき、日本語キーワード語句（３１）として使用者が形容詞や助詞を含めた場合や、複数のキーワード語句を１個のキーワード語句として入力した場合には、周知の処理方法によって基本形に変形したり、分割して複数のキーワード語句にしてもよい。この際、形態素解析等の言語処理方法が用いられることは公知である。

もっとも本発明において対応語句記憶部（６０）が最適な対応語句を抽出する上で、助詞や形容詞が重要な働きを果たす場合が多く、なるべくそれらを含めた形で対訳コーパスデータベース（５２）から対訳文を抽出するのが望ましい。助詞は後述する係り受け関係を特定するのに有効であるし、形容詞が含まれることで対応語句の多義性の解消などに寄与すること

も考えられる。

また、上記対訳文抽出の際に、入力したキーワードに対応する対訳文が対訳コーパスデータベース（５２）に見つけられない場合には、再び入力部（４０）に処理を戻して、使用者に再入力を求めるようにしてもよい。或いは、シソーラスを用いて自動的に他のキーワード語句に置き換えるように構成してもよい。

具体的には処理部（５１）に図示しないキーワード整形部を設け、入力部（４０）で入力されたキーワード語句を、整形処理する。該処理では、キーワード語句を公知の形態素解析処理により形態素に分割し、キーワード語句が複数の形態素から成る場合には、上記コーパスにおける接続助詞や格助詞を適宜削除したり、或いは対訳コーパス中に存在する形に合わせて加えたりする。形容詞に含む語尾、例えば「否定的な」の「な」を削除・変形させてもよい。

また、記憶手段にシソーラスを格納した上で、該キーワード語句の全形態素又は一部形態素を置換してもよい。

次の対応語句記憶部（６０）では、対訳文抽出部（５０）で抽出された対訳文から、日本語のキーワード語句を含む日本語対応語句に対応する英語対応語句を、部分対応情報に基づいて抽出し、対応語句群テーブル（５３）として記憶手段に記憶する。

すなわち、図７の例では「to the park」「I went ...」「with her」が記憶される。

次に、以上により形成された対応語句群テーブル（５３）を、図１０に示すテキスト生成部（７０）に入力し、英語テキストを生成する。

いくつかの語句を入力し、その語句を含むテキストを生成する方法としては次のような手法がある。すなわち、本件出願人が前記の特許文献３で開示するテキスト生成方法を、翻訳先言語である英語に適用して用いる。

本テキスト生成部（７０）の具体的な構成例として図１０に示す各部を備える。テキスト生成部（７０）は、例えばＣＰＵとメモリ、ハードディスクなどの外部記憶媒体を備えるパーソナルコンピュータなどにより構成

することができ、主な処理をＣＰＵにおいて行い、処理の結果を随時メモリ、外部記憶媒体に記録する。

本実施例で、入力された英語対応語句が、単語列ではなく単語列の主辞となる内容語である場合には、テキスト候補生成部（７３）における処理の前に、単語列の候補を生成する。これは英語対応語句が内容語だけの場合、テキスト候補生成部（７３）において係り受け関係を決定しただけではテキストが形成されない場合があるからである。

該処理において、入力された英語対応語句（５３）は２つの処理に用いられる。その１つは単語列生成規則獲得部（７１）であり、もう１つは単語列候補生成部（７２）である。以下では英語対応語句（５３）のうち、単語列の主辞となる内容語であるものを特に英語対応単語と呼び、英語対応語句（５３）が英語対応単語である場合には単語列候補生成部（７２）で処理する一方、該当しない場合にはテキスト候補生成部（７３）に英語対応語句（５３）を送る。

内容語は、その語の品詞が、動詞、形容詞、名詞、指示詞、副詞、接続詞、連体詞、感動詞、未定義語である形態素の見出し語であるとし、それ以外の形態素の見出し語を機能語とする。

単語列生成規則獲得部（７１）では、英語対応単語が与えられたとき、それぞれを含む文を対訳コーパス（７５）から検索し、形態素解析、構文解析(係り受け解析）をする。

そして、そこから英語対応単語を含む単語列を抽出して、英語対応単語から英語対応語句（５３）を生成する単語列生成規則（７６）を獲得し、記録する。このとき、対訳コーパス（７５）を用いて、英語と日本語の対応付けをした単語列生成規則とするので、上記英語対応単語に対応する日本語の単語も同時に単語列として生成することができる。

例えば、「1995」→「before the end of 1995／一九九五年中の」、「possibility」→「to the possibility／可能性に」などの単語列生成規則（７６）を獲得し、記録する。

なお、ここでは英語対応単語に着目して英語と日本語の対応語句の組を

生成したが、日本語キーワードから英語と日本語の対応語句の組を生成することも可能である。

ここで、生成規則の自動獲得には次の手法を用いる。英語対応語句の集合をVとし、英語対応語句k(∈V)から単語列を生成する規則の集合をRkとするとき、規則rk(∈Rk)は次の形式で表現されるものと定義する。

k → hk m*

hkは英語対応語句を含む主辞形態素、m*は同じ単語列内でhkに連続する任意個の形態素とする。英語対応単語が与えられると、この形式を満たす規則を翻訳先言語のコーパス（７５）から自動獲得する。

一方、単語列候補生成部（７２）では、単語列生成規則（７６）を参照しながら、入力された英語対応語句（５３）から出力する英語テキスト（３２）を構成する単語列の候補を生成する。日本語テキストも同時に出力する場合には、このときに合わせて日本語対応語句についても単語列の候補を生成する。

例えば、「1995」では自然なテキストを構成する単語列とはなりにくいが、「before the end of 1995」あるいは「in 1995」のように「1995」という単語と極めて密接な関連性を有する語句を付加し、後段の処理によるテキスト生成に備える。

本実施例のように、単語列生成規則獲得部（７１）により対訳コーパス（７５）から入力する英語対応語句（５３）（及び日本語対応語句）の単語列規則を生成することで、最小限の計算量で効果的に単語列生成規則を得ることができ、処理速度の向上に寄与する。

もっとも、必ずしも英語対応語句（５３）に関連する語句をコーパスから抽出する構成を取る必要はなく、計算能力に応じて任意の語句を入力された英語対応語句（５３）の前後に付加してもよい。あるいは、別に対訳辞書データベースを備えて、それに含まれる慣用表現の情報から単語列を生成することもできる。上記「possibility」→「to the possibility」などは対訳辞書データベースに記載される表現であり、単語列の候補として生成することができる。

また、日本語など主格を多く省略する言語を入力した場合には、「respond」→「He responded」などのように主語を補って単語列候補を生成することができる。このとき、日本語などの多くの言語では主格が明らかな時や、形式主語であるときに省略されることに着目し、入力に主格が何であるかの情報だけでなく、主格がないという情報を用いることで、「respond」→「He responded」を生成せず、「respond」→「It is responded that」を生成するようにすることもできる。

次に、テキスト候補生成部（７３）においてテキスト候補を生成する。テキスト候補はグラフあるいは木の形で表現する。ここでは英語対応語句（５３）のうち、「to the park」「I went ...」「with her」の３語句の関係を例として説述する。

すなわち、図１１のように、各英語対応語句（５３ａ）（５３ｂ）（５３ｃ）の間に係り受けの関係を仮定して、テキスト候補（５４）のような英語対応語句を単位とした依存構造木の形でテキスト候補を生成する。このとき、３語の場合に全ての係り受け関係は３！×２＝１２通りであるが、翻訳先言語の文法・特性に合わせて語順の固定などにより候補の数を削減することができる。

生成されたテキスト候補（５４）は、評価部（７４）でコーパスから学習した英語対応語句生成モデル（７７）や言語モデル（７８）を用いて順序付けされる。以下、英語対応語句生成モデル（７７）と、言語モデル（７８）として形態素モデル及び係り受けモデルについて説述する。

英語対応語句生成モデルでは、次の５種類の情報を素性として用いたモデル（ＫＭ１ないし５）を考える。以下で、英語対応語句の集合Ｖは、ある回数以上コーパスに出現した主辞単語の集合とし、単語列は前記で表現されるものと仮定する。また、各英語対応語句は独立であり、与えられたテキストが単語列 w1・・・wm からなるとき、英語対応語句 ki は単語 wj (1≦j≦m) に対応していると仮定する。図１２にモデルの説明図を示す。

[ＫＭ１]

前方の二単語を考慮(trigram)

ki は前方の二単語 wj-1 と wj-2 のみに依存すると仮定する。

$$P(K|M,D,T) = \prod_{i=1}^{n} P(k_i|w_{j-1},w_{j-2})$$

[KM2]

後方の二単語を考慮(後方 trigram)

ki は後方の二単語 wj+1 と wj+2 のみに依存すると仮定する。

$$P(K|M,D,T) = \prod_{i=1}^{n} P(k_i|w_{j+1},w_{j+2})$$

[KM3]

係り単語列を考慮(係り単語列)

ki を含む単語列に係る単語列がある場合、ki はそのうち最も文末側の単語列の末尾から二単語 wl と wl-1 のみに依存すると仮定する(図１２参照)。

$$P(K|M,D,T) = \prod_{i=1}^{n} P(k_i|w_l,w_{l-1})$$

[KM4]

受け単語列を考慮(受け単語列)

ki を含む単語列を受ける単語列がある場合、ki はその単語列内の主辞単語から二単語 ws と ws+1 のみに依存すると仮定する(図１２参照)。

$$P(K|M,D,T) = \prod_{i=1}^{n} P(k_i|w_s,w_{s+1})$$

[KM5]

係り単語列を最大二単語列考慮(係り二単語列)

ki を含む単語列に係る単語列がある場合、ki は、そのうち最も文末側の単語列の末尾から二単語 wl 、wl-1 と、最も文頭側の単語列の末尾から二単語 wh 、wh-1 のみに依存すると仮定する(図１２参照)。

$$P(K|M,D,T) = \prod_{i=1}^{n} P(k_i|w_l,w_{l-1},w_h,w_{h-1})$$

次に、形態素モデル（MM）について示す。形態素に付与するべき文法

的属性がｌ個あると仮定する。テキストつまり文字列が与えられたとき、その文字列が形態素であり、かつ ｊ(1≦ｊ≦ｌ) 番目の文法的属性を持つとしたときの尤もらしさを確率値として求めるモデルを用いる。

テキストＴが与えられたとき、順序付き形態素集合Ｍが得られる確率は、各形態素ｍｉ(1≦ｉ≦ｎ) が独立であると仮定し、

$$P(M|T) = \prod_{i=1}^{n} P(m_i|m_1^{i-1}, T)$$

と表す。ここで、ｍｉ はｌからｌまでのいずれかの文法的属性を表わす。

一方、係り受けモデル（ＤＭ）は、テキストＴと順序付き形態素集合Ｍが与えられたとき、各単語列に対する係り受けの順序付き集合Ｄが得られる確率は、各々の係り受けｄｌ・・・ｄｎ が独立であると仮定し、

$$P(D|M,T) = \prod_{i=1}^{n} P(d_i|M,T)$$

と表わす。

例えば、「to the park」「I went ...」「with her」の３つの英語対応語句（５３）から「I went with her to the park.」と「I went to the park with her」の２つの候補が生成されたとする。係り受けモデルにより、このうち尤もらしい係り受け構造を持つ候補が優先される。

以上に示すような各モデルを用い、本発明では評価部（７４）においてテキスト候補（５４）に評価付けを行う。

評価部（７４）では上記手法により句と句の依存関係や、形態素の並びとしての尤もらしさなどが考慮されるため、例えば英語における３単現のｓの有無などについても、適切なものが評価値が高くなるので、文法的な正確さにも寄与する。

そして、評価値が最大あるいは閾値を超えるテキスト候補、あるいは評価値の上位Ｎ個を表層文に変換して出力する。

出力部（８０）における出力方法としては、モニタによる表示の他、音声合成を用いた発声、翻訳システムなど他の言語処理システムへのデータ

出力などが可能である。また、ネットワーク接続された他のコンピュータなどにテキストデータを送出してもよい。

本発明は、以上のように英語テキスト（３２）を生成するものであるが、最後に文法的な補正処理を加えてもよい。すなわち、上記のように文法的にもある程度正しい出力が可能であるが、本方法による生成では時制の誤りや前置詞・主語の欠落などが生じる可能性もある。その場合、公知のＯＣＲ（光学的文字読み取り認識）技術における誤り修正の手法を適用することが考えられる。

英語側のテンス(時制)、(相：完了形、進行形などで表わされる)、モダリティ(法相：may, can, must などで表わされる)に不整合がある場合は、本件出願人らによる特許第３３８８３９３号公報に開示した方法などにより修正することができる。

例えば、「彼女と公園に行った」なら時制が過去と推定して、英語でも過去形を用いる、「彼女と公園に行ってきたところだ」なら完了形を用いる、「彼女と公園に行くだろう」なら、英語で may を用いる、というように間違った英語が選択された場合に修正する。

また、三単現のｓや前置詞の間違いなどは、例えば、下記文献１に開示されるような文法的誤りのパターンを機械学習させ、誤りの検出を行う手法などにより、修正することができる。

〔文献１〕

Emi Izumi, Kiyotaka Uchimoto, Toyomi Saiga, Thepchai Supnithi, Hitoshi Isahara,「Automatic Error Detection in the Japanese Learners' English Spoken Data」, Proceedings of the ACL2003 Interactive Poster/Demo Sessions pp.145-148, 2003

（実施例２）

本発明の実施例２として、図３にフローチャートを示す処理がある。すなわち、日本語キーワード語句（１）を入力して対訳文を抽出（２）した際、複数の対訳文が抽出され、その部分対応情報から日本語対応語句が複

数の種類存在する場合に、日本語キーワード語句の絞り込み処理（２０）を行うようにする。

図９に従って説述すると、対訳文抽出部（５０）で日本語キーワード（３１）を含む対訳文を対訳コーパスデータベース（５２）から抽出する。例えばキーワード語句として「彼女」を入力したとき、複数の対訳文中に「彼女が」「彼女と」「彼女に」が日本語対応語句として抽出されることがある。

本実施例に係る日本語語句候補提示部（６１）は、これらの日本語対応語句を使用者にすべて提示し、使用者はいずれの日本語対応語句がキーワード語句として最適であるか選択するようにする。

選択にはマウス（６２）、キーボード（６３）などを用い、使用者への提示はモニタ（６４）で表示する。また、タッチパネルモニタ（６５）を用いて優れたユーザインタフェースを提供することもできる。

本実施例では、同様に「公園」と入力した場合には「公園へ／to the park」「公園で／in the park」を、「行く」の場合には「行く／I will go」「行った／I went ...」などを候補とし提示する。このように使用者がキーワード語句を入力するたびに対訳コーパスデータベース（１０）から選択できる対応語句を提示することで、使用者の介入を容易にしながら、より適切なテキスト生成を図るようにする。

さらに、周知の文字入力方法としてローマ字や読み仮名の最初の１文字を入力した時点から順にその文字から始まる単語列を表示する手法がある。これを本実施例に適用すると、例えばｋと入力した時点で「彼は」「彼女は」「今日」・・・などが表示され、kan まで入力すると「彼女と」「彼女が」・・と絞られるようになる。対訳コーパスデータベースからこれらの候補を漸次抽出するのが処理上困難である場合には、適当な辞書データベースを別に設けて該辞書で単語のレベルまで絞りをかけた後に、対訳コーパスデータベースから日本語対応語句を抽出すると良い。

（実施例３）

本発明の実施例３として、図４にフローチャートを示す処理がある。ここでは、対訳文を抽出（２）した際に、該対訳文においてキーワード語句と共起する語句を抽出（２１）する。抽出された共起語句は使用者に提示（２２）し、使用者が選択した共起語句は新たなキーワード語句（１）として追加する。

図１３に示すように、日本語キーワード語句で「彼女」「公園」と入力した時点で、対訳文抽出部（５０）が「彼女と公園へ行った／I went to the park with her」を抽出し、共起語句抽出部（６６）は「彼女と」「公園へ」と共起する語句として「行った」を抽出する。このような共起語句の抽出方法は公知である。

そして、共起語句提示部（６７）でモニタ（６４）等から使用者に対して「行った」を提示し、使用者がそれをキーワード語句とするのが適当と判断した場合にはマウス（６２）（６３）から選択することによりこれを新たなキーワードとして再び対訳文抽出部（５０）に入力するか、対応語句記憶部（６０）において「公園へ／to the park」を対応語句群テーブル（５３）に記憶する。

前者の場合にはさらに選択した共起語句と共起する語句を選択することができるが、対訳文の数が膨大になる可能性があるため、後者の方法でもよい。

（実施例４）

本発明の実施例４に係る構成は、図５に示すように、日本語キーワード語句を入力すると、同時に２つの言語についてテキスト生成を行うテキスト生成方法である。

すなわち、図示の例では日英対訳コーパスデータベース（１０ａ）と日本語対語対訳コーパスデータベース（１０ｂ）を用いてそれぞれについて対訳文抽出（２ａ）（２ｂ）、部分対応情報（１１ａ）（１１ｂ）を用いた対訳語句記憶（３ａ）（３ｂ）、得られた対応語句群記憶テーブル（１２ａ）

（１２ｂ）からテキスト候補生成（４ａ）（４ｂ）、評価（５ａ）（５ｂ）を行い、英語テキスト（６ａ）、タイ語テキスト（６ｂ）を同時に出力する。

これらの各方法において、上記実施例１ないし３で述べたような処理方法を導入してもよい。

本構成では、複数の言語テキストを同時に出力できるため、ネットワーク上において複数の言語の使用者が共存する場合などに特に好適である。

（実施例５）

実施例５は、テキスト候補生成において、日本語テキスト候補と英語テキスト候補を同時に生成し、使用者に生成された他言語の内容把握を容易にするものである。

図６に示すように、対応語句を記憶（３）する際に、対応語句群テーブル（１２）に日英の対訳語句を共に記憶しておき、英語テキスト候補生成（４）に合わせて日本語テキスト候補を生成（２３）する。両言語における係り受け関係を対応させておくことにより、生成された両テキストは同内容の対訳テキストが得られていると考えられるため、これらを使用者に提示することで、使用者は日本語による生成内容の確認を行うことができる。

また、日本語テキスト候補の中から適切な係り受け関係になっているものを使用者が選択するようにすることで、係り受け関係を特定することができるため、英語テキスト候補の中から、係り受け関係が正しくかつ自然なテキストを得ることができる。

（実施例６）

以上説述した実施例１ないし５は、いずれも日本語キーワード語句を直接入力するものであったが、本発明を次のようなシステムに実装して利用することができる。すなわち、本システムでは図１４に示すように、ユーザーは日本語テキストを入力する。例えば、「彼女は公園へ行った」と入力部（４０’）（前記入力部（４０）と入力する対象がテキストである他は同様の構成である）で入力すると、次のようなキーワードの抽出処理を図１

５の構成図におけるキーワード抽出部（９０）で行う。

キーワード抽出部（９０）の構成を図５に示す。ここでもＣＰＵ及びメモリが協働して各処理を行う。キーワード抽出部（９０）では、入力された日本語入力テキストからそのテキストの内容を特徴的に表すキーワードを抽出する。

このような技術は、言語処理において文書を要約する技術や、文書検索などの要素技術として公知の多数の手法が知られており、それらを適宜用いることができるが、ここでは一例として下記文献２に記載の方法を用いる。

〔文献２〕

久光徹、丹羽芳樹、辻井潤一，「タームの representativeness」を測る」，情報処理学会自然言語処理研究会　１９９９－ＮＬ－１３３，１９９９

本方法によると、特徴語を選ぶために文書中の単語の話題性もしくは分野代表性（representativeness、本明細書ではこれを特徴性と呼ぶ。）を測ることが可能であり、かつ数値的な評価によるため、本発明の実施に好適である。以下に、簡単に説述する。

まず、キーワード抽出部（９０）では、公知の形態素解析技術を用いて、日本語テキストを形態素解析部（９１）において形態素解析する。解析された形態素はメモリ又は図示しない外部記憶装置などに形態素テーブルとして記録する。

そして、形態素テーブルから形態素を順次読み出し、その形態素（以下、これを着目タームと呼ぶ）毎に特徴性を測る。

まず文書抽出部（９２）において、着目タームWについて、Wを含む文書すべてを任意の文書データベース（９３）から抽出する。文書データベース（９３）は複数の日本語（翻訳元言語）の文書が含まれたものであり、外部記憶装置などに記憶されている。日本語単言語のコーパスや日英の対訳コーパスの日本語部分を用いてもよい。

次に、着目タームWが抽出された文書すべての集合における単語分布と

、文書データベース（９３）に含まれる全文書の単語分布とを、単語分布算出部（９４）において算出し、各単語分布間の異なりの度合いを測る。

具体的には異なり度合算出部（９５）において次のような計算処理を行う。

すなわち、着目タームW、Wを含む文書すべての集合D（W）、全文書の集合D_0、D（W）における単語分布$P_{D(W)}$、D_0における単語分布P_0として、Wの特徴性Ｒｅｐ（W）を、２つの分布｛$P_{D(W)}$，P_0｝の距離Dist｛$P_{D(W)}$，P_0｝に基づいて定義する。

単語分布間の距離計測方法として、本実施例では対数尤度比を用いている。すなわち、全単語を｛$W_1, \cdots, W_n$｝、単語w_iがD（W）、D_0に出現する頻度をそれぞれk_i、K_iとするとき、$P_{D(W)}$、P_0の距離Dist｛$P_{D(W)}$，P_0｝を、次のように定義する。

$$Dist(P_{D(W)}, P_0) = \sum_{i=1}^{n} k_i \log \frac{k_i}{\#D(W)} - \sum_{i=1}^{n} k_i \log \frac{K_i}{\#D_0}.$$

ここで、＃D（W）は着目タームWについてD（W）の含む単語数、＃D_0は同様に全文書の含む単語数である。この数式の定義によると、＃D（W）が離れた着目ターム同士の特徴性を有効に比較することが難しいため、次の数式のように正規化を行った特徴性Ｒｅｐ（W）を定義する。なおＢ（・）は＃D（W）が適切な数となる範囲内（例えば１０００≦＃D（W）≦２００００）で特徴性が精度よく求められるような指数関数を用いた近似関数である。

Ｒｅｐ（W）＝Dist｛$P_{D(W)}$，P_0｝／Ｂ（＃D（W））

ここで、「する」などのように著しく＃D（W）が大きい場合には、D（W）の抽出数を限定し、＃D（W）≦２００００を満たすようにすることで、上記近似関数を有効に用いることができると共に計算量を削減できる。

キーワード抽出部（９０）では以上の方法により特徴性を算出すると共に、所定の閾値に従って、キーワード決定部（９６）により入力した日本語入力テキストのキーワードを抽出する。

ここで、例えば「彼女」「公園」「行く」がキーワードとして抽出されるので、上記実施例と同様に、対訳文抽出部（５０'）により対訳コーパスデータベース（１０）から対訳文を抽出する。上記実施例では説明のため省略したが、このとき例えば「彼女は動物園へ行った。／She went to the zoo.」なども同時に抽出されている。

そして、対訳語句記憶部（６０'）も同様であり、テキスト生成部（７０'）に進む。

以上、各処理部（４０'）（５０'）（６０'）（７０'）は前記実施例の（４０）（５０）（６０）（７０）と同態様の処理部であって、特記しない構成は同一である。

前記実施例のテキスト生成部（７０）は図１０に示すような構成であったが、ここでは例えば評価部（７４）で閾値を超えるテキスト候補を、実施例５のように複数の対訳文の形で出力し、最後に類似度評価部（１００）において、対訳文のうち日本語テキストと、最初に入力した日本語入力テキストの類似度を評価する。

類似度の評価方法としては、例えばテキストに含まれる文字列の一致する割合がどの程度であるかを算出して求める方法、あるいは下記文献３に開示されるような自動翻訳した結果と人間の翻訳結果を文字列の単位（或いは単語単位）で比較してその一致度を基に計算する方法などを用いることができる。

〔文献３〕

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu,「Bleu:a Method for Automatic Evaluation of Machine Translation」, IBM Research Report RC22176(W0109-022) 2001

類似度評価部（１００）では、「彼女は公園へ行った」という入力テキストと、テキスト生成部（７０'）で生成された「彼女と公園へ行った」「彼女は公園へ行った」の類似度を比較し、より類似度の高い「彼女は公園へ行った。／She went to the park」の対訳文を出力部（８０'）から出力することができる。

以上、本発明の実施例１～実施例６まで説述した。上記では説明の便宜のために、各部（４０）（５０）（６０）（７０）（８０）を別個に説述したが、これらは一体的に例えば１台のパーソナルコンピュータによって提供することができる。特に、ＣＰＵ、メモリ、入出力装置、ネットワークに接続するためのネットワークアダプタ（図示していない）、外部記憶装置などは共用することが望ましく、装置の簡略化に寄与することができる。

外部記憶装置に記録される対訳コーパスデータベース（１０）、コーパス（７５）はいずれも同一のデータベースの一部又は全部を用いることが可能である。

また、これらは外部記憶装置上に記録される場合にとどまらず、ネットワーク上の複数のサーバーに記録されたものを収集するように構成してもよい。

以上詳述したように、本発明は、キーワード語句を与えることによって、対訳コーパスから他言語のテキストを生成することができるので、自然な他言語を出力することができる。また、キーワードを入力することにより、原言語がテキストである場合に比して処理が容易であると共に、原言語テキストの解析誤りによる他言語テキストの誤りがなく、より正確なニュアンスのテキスト生成に寄与する他言語テキスト生成方法又は他言語テキスト生成装置を提供することができる。

加えて、使用者に対してキーワード語句を提示することにより、使用者においては原言語で提示されるために理解が容易で指示が簡便に行える一方、本方法を備えた装置では正確なキーワード語句を用いて処理が行えるため、高精度な他言語テキストの生成が可能な他言語テキスト生成方法又は他言語テキスト生成装置を提供することができる。

さらに、対訳文抽出の際に、キーワード語句を変形させることにより、効率的な対訳文抽出処理が行える。この際、複数の形態素から成る場合には例えば語尾の助詞を削除したり、変形させたりして、対訳コーパス中に完全に一致するキーワード語句がなくとも抽出が行えるようにする。また、同義語、狭義語、広義語などの類語に置き換えることも可能な他言語テキスト生成方法又は他言語テキスト生成装置を提供することができる。

また、原言語１言語のキーワード語句を入力するだけで、同時に複数の言語のテキストを生成することができるので、効率の向上が図れるだけでなく、同時に多くの言語の使用者とのコミュニケーションにも寄与する他言語テキスト生成方法又は他言語テキスト生成装置を提供することができる。

さらに、原言語のテキスト候補を他言語のテキスト候補と共に出力することができるので、使用者が生成された他言語のテキストの意味を正確に把握することが可能な他言語テキスト生成方法又は他言語テキスト生成装置を提供することができる。

また、評価する処理を行うことにより、テキスト候補が複数ある場合にも、自動的に１個又は特定の候補数だけテキストを出力することが可能な他言語テキスト生成方法又は他言語テキスト生成装置を提供することができる。これによって、例えば、後述の学習モデルによる確率値に応じて確率の高いものから所定数だけ順序付けして出力することもできる。

産業上の利用可能性

母国語などのキーワード語句をいくつか入力することで、そのキーワード語句の対訳語句を用いる他言語のテキストを生成するので、翻訳技術の向上に寄与することができ、産業上の利用価値が高い。

請求の範囲

1. 原言語の語句をキーワード語句として入力することにより、原言語とは異なる他言語のテキストを生成する他言語テキスト生成方法であって、入力手段から、単数又は複数の該原言語のキーワード語句を入力する入力ステップ、対訳文中の語句間対訳関係に係る部分対応情報を含む原言語・他言語間の対訳コーパスデータベースを用い、対訳文抽出手段が、該キーワード語句を含む対訳文を、該対訳コーパスデータベースから抽出する対訳文抽出ステップ、該対訳文の部分対応情報から、各原言語のキーワード語句を含む原言語対応語句に対応する他言語の各他言語対応語句で構成する対応語句群テーブルを記憶手段に記憶する対応語句記憶ステップ、テキスト候補生成手段が、該対応語句群テーブルに含まれる各他言語対応語句間の係り受け関係を仮定して他言語のテキスト候補を生成するテキスト候補生成ステップ、出力手段から、少なくとも１つのテキスト候補を出力する出力ステップの各ステップを含むことを特徴とする他言語テキスト生成方法。

2. 前記他言語テキスト生成方法の対訳文抽出ステップにおいて、前記入力ステップで入力したキーワード語句に対して、複数の対訳文が抽出され、部分対応情報から原言語対応語句が複数の種類存在するときに、該対訳文抽出ステップの次に、複数の原言語対応語句を使用者に選択可能に提示する原言語語句候補提示ステップを備え、対応語句記憶ステップにおいて、使用者が選択した場合に、その原言語対応語句に対応する他言語対応語句を対応語句群記憶テーブルに記憶することを特徴とする請求項１に記載の他言語テキスト生成方法。

3. 前記入力ステップにおいて、１個のキーワード語句を入力する毎に、前記対訳文抽出ステップ及び、前記対応語句記憶ステップの各処理を行うと共に、抽出された対訳文中において該キーワード語句と共起

する共起語句を抽出し共起語句テーブルに記憶する共起語句抽出ステップ、該共起語句テーブル中の共起語句を使用者に選択可能に提示する共起語句提示ステップの各ステップを備え、該入力ステップにおいて、使用者が共起語句を選択した場合には、該共起語句を新たなキーワード語句として該入力ステップにおいて入力し、全てのキーワード語句の入力が終了した後に、前記テキスト候補生成ステップに進むことを特徴とする請求項１又は２に記載の他言語テキスト生成方法。

４．　前記他言語テキスト生成方法の対訳文抽出ステップにおいて、各処理に先だち、入力ステップで入力されたキーワード語句について、構成する一部の形態素の加減、又は類語への置換を行うことを特徴とする請求項１ないし３に記載の他言語テキスト生成方法。

５．　前記他言語テキスト生成方法において、他言語テキストが複数の言語であって、対訳文抽出ステップ、対応語句記憶ステップ、テキスト候補生成ステップにおいて、前記原言語と、各他言語との間についてそれぞれ処理を行い、出力ステップにおいて、複数の言語のテキスト候補を出力することを特徴とする請求項１ないし４に記載の他言語テキスト生成方法。

６．　前記他言語テキスト生成方法のテキスト候補生成ステップにおいて、テキスト候補生成手段が、該対応語句群テーブルに含まれる各他言語対応語句間の係り受け関係を仮定して他言語のテキスト候補を生成すると共に、原言語テキスト候補生成手段が、該対応語句群テーブルに含まれる各原言語対応語句間の係り受け関係を仮定して原言語のテキスト候補を生成し、出力ステップにおいて、出力手段から、少なくとも１組の原言語及び他言語の対訳テキスト候補を共に出力することを特徴とする請求項１ないし５に記載の他言語テキスト生成方法。

7．　前記他言語テキスト生成方法において、テキスト候補生成ステップの次に、評価手段が、該テキスト候補を評価付けする評価ステップを有し、出力ステップにおいては、該評価に基づいて少なくとも１つのテキスト候補を出力することを特徴とする請求項１ないし６に記載の他言語テキスト生成方法。

8．　原言語の単語をキーワードとして入力することにより、原言語とは異なる他言語のテキストを生成する他言語テキスト生成装置であって、単数又は複数の該原言語のキーワード語句を入力する入力手段と、対訳文中の語句間対訳関係に係る部分対応情報を含む原言語・他言語間の対訳コーパスデータベースと、該キーワード語句を含む対訳文を、該対訳コーパスデータベースから抽出する対訳文抽出手段と、該対訳文の部分対応情報から、各原言語のキーワード語句を含む原言語対応語句に対応する他言語の各他言語対応語句で構成する対応語句群テーブルを記憶可能な対応語句記憶手段と、該対応語句群テーブルに含まれる各他言語対応語句間の係り受け関係を仮定して他言語のテキスト候補を生成するテキスト候補生成手段と、少なくとも１つのテキスト候補を出力する出力手段とを少なくとも備えたことを特徴とする他言語テキスト生成装置。

9．　前記他言語テキスト生成装置が、入力したキーワード語句に対して前記対訳文抽出手段により複数の対訳文が抽出され、その部分対応情報から原言語対応語句が複数の種類存在するか否か判定し、複数の種類存在する場合には、使用者に該各原言語対応語句を提示する原言語語句候補提示手段を備えると共に、前記入力手段から、使用者が提示された原言語対応語句の１個を選択可能であり、使用者が選択した場合には、前記対応語句記憶手段がその原言語対応語句に対応する他言語対応語句を対応語句群記憶テーブルに記憶する請求項８に記載の他言語テキスト生成装置。

１０．　前記他言語テキスト生成装置が、入力手段から１個のキーワード語句を入力する毎に、前記対訳文抽出手段及び、前記対応語句記憶手段が作用する構成において、抽出された対訳文中において該キーワード語句と共起する共起語句を抽出し共起語句テーブルに記憶する共起語句抽出手段と、該共起語句テーブル中の共起語句を使用者に選択可能に提示する共起語句提示手段とを備え、該入力手段から使用者が共起語句を選択した場合には、該共起語句を新たなキーワード語句として入力し、全てのキーワード語句の入力が終了した後に、前記テキスト候補生成手段が作用することを特徴とする請求項８又は９に記載の他言語テキスト生成装置。

１１．　前記他言語テキスト生成装置において、前記入力手段から入力されたキーワード語句について、構成する一部の形態素の加減、又は類語への置換を行うキーワード整形手段を備え、対訳文抽出手段において処理を行うことを特徴とする請求項８ないし１０に記載の他言語テキスト生成方法。

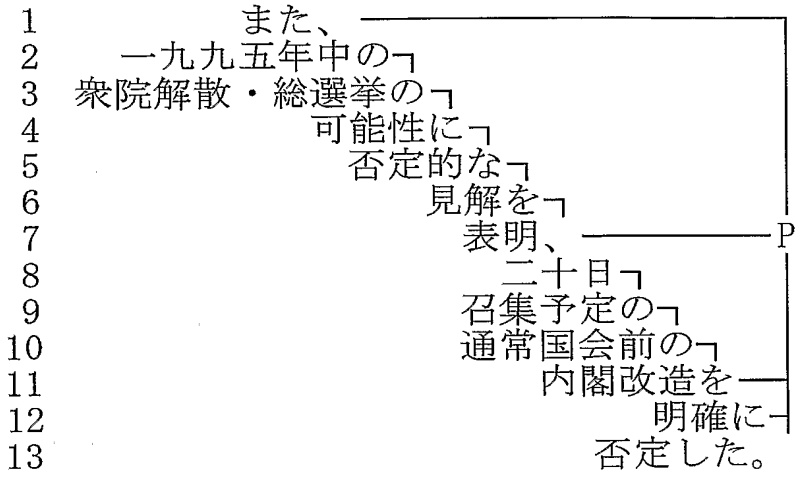
１２．　前記他言語テキスト生成装置において、対訳コーパスデータベースに、原言語と、複数の他言語との間の対訳文中の語句間対訳関係に係る部分対応情報を含み、対訳文抽出手段と、対応語句記憶手段と、テキスト候補生成手段において、該原言語と、各他言語との間についてそれぞれ処理を行うと共に、出力手段から、複数の言語のテキスト候補を出力することを特徴とする請求項８ないし１１に記載の他言語テキスト生成装置。

１３．　前記他言語テキスト生成装置において、前記テキスト候補生成手段が、前記対応語句群テーブルに含まれる各他言語対応語句間の係り受け関係を仮定して他言語のテキスト候補を生成すると共に、該対応語句群テーブルに含まれる各原言語対応語句間の係り受け関係を仮定して原言語のテキスト候補を生成する原言語テキスト候補生成手

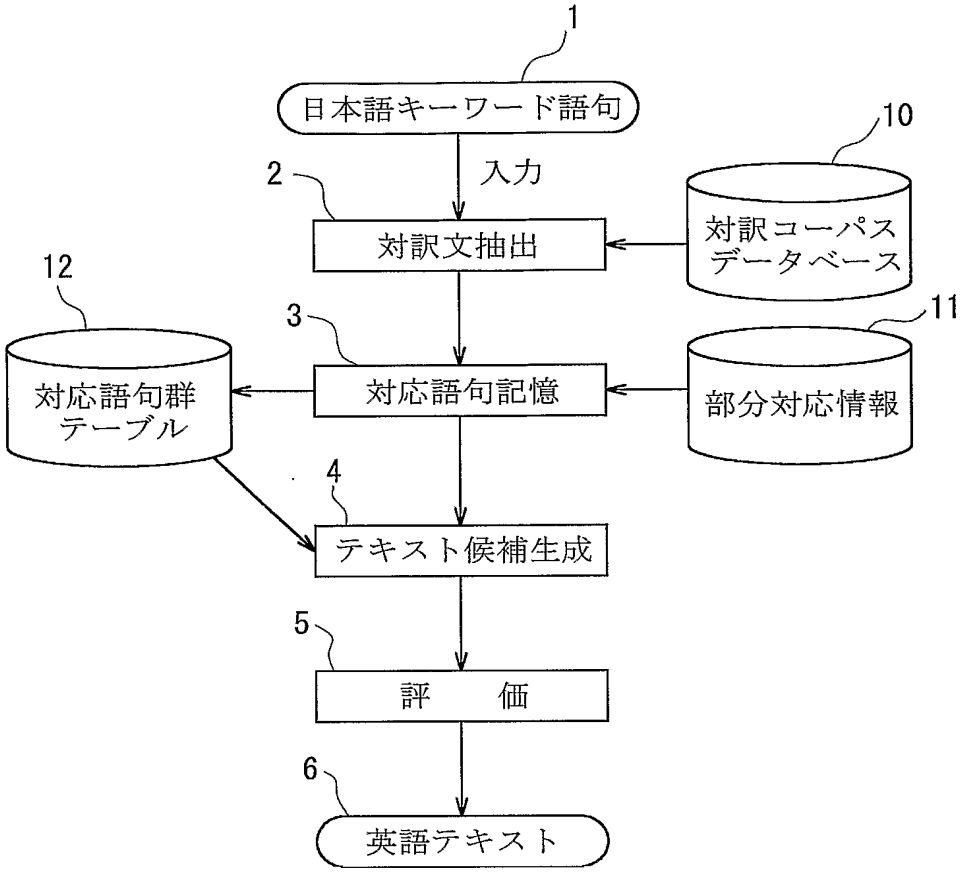
段を備え、出力手段から、少なくとも１組の原言語及び他言語の対訳テキスト候補を共に出力することを特徴とする請求項８ないし１２に記載の他言語テキスト生成装置。

１４．　前記他言語テキスト生成装置において、前記テキスト候補を評価付けする評価手段を備えたことを特徴とする請求項８ないし１３に記載の他言語テキスト生成装置。

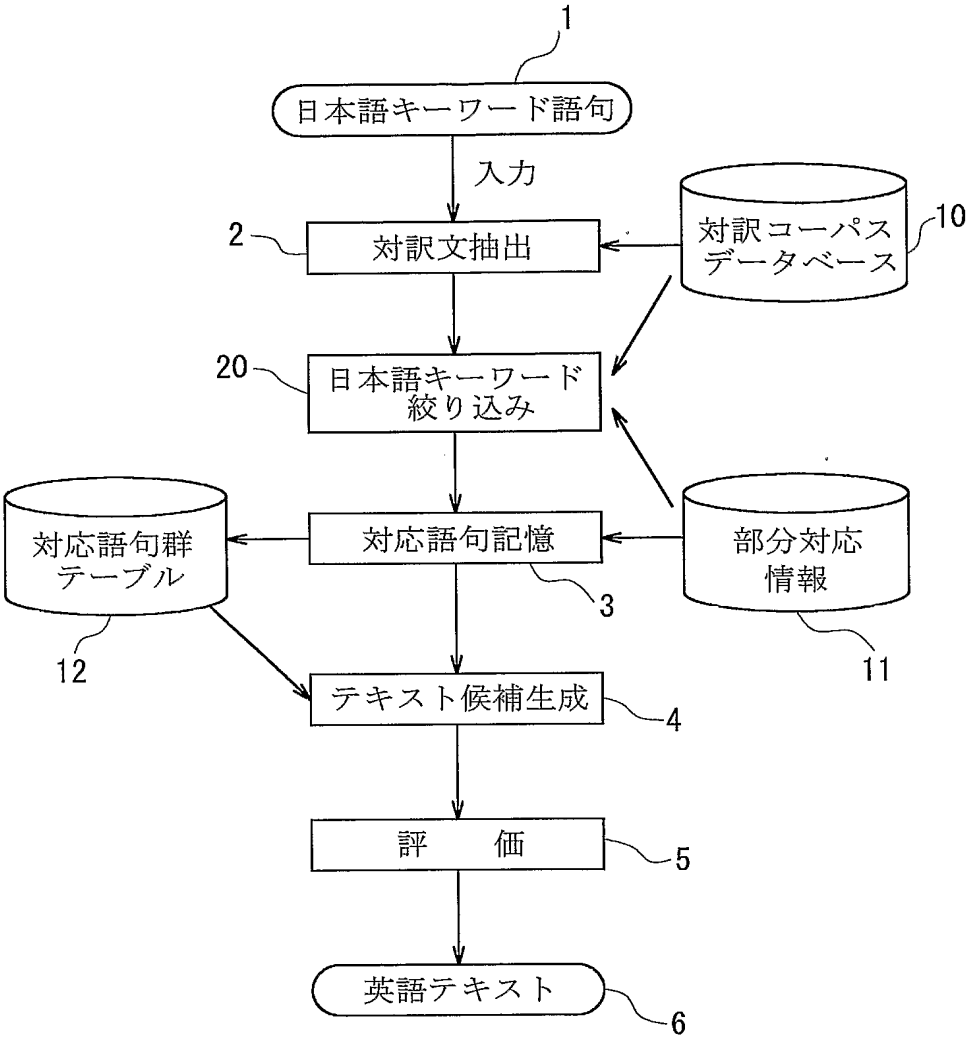
第１図


1
2
3
4
5
6
7
8
9
10
11
12
13
また、
一九九五年中の
衆院解散・総選挙の
可能性に
否定的な
見解を
表明、
P
二十日
召集予定の
通常国会前の
内閣改造を
明確に
否定した。

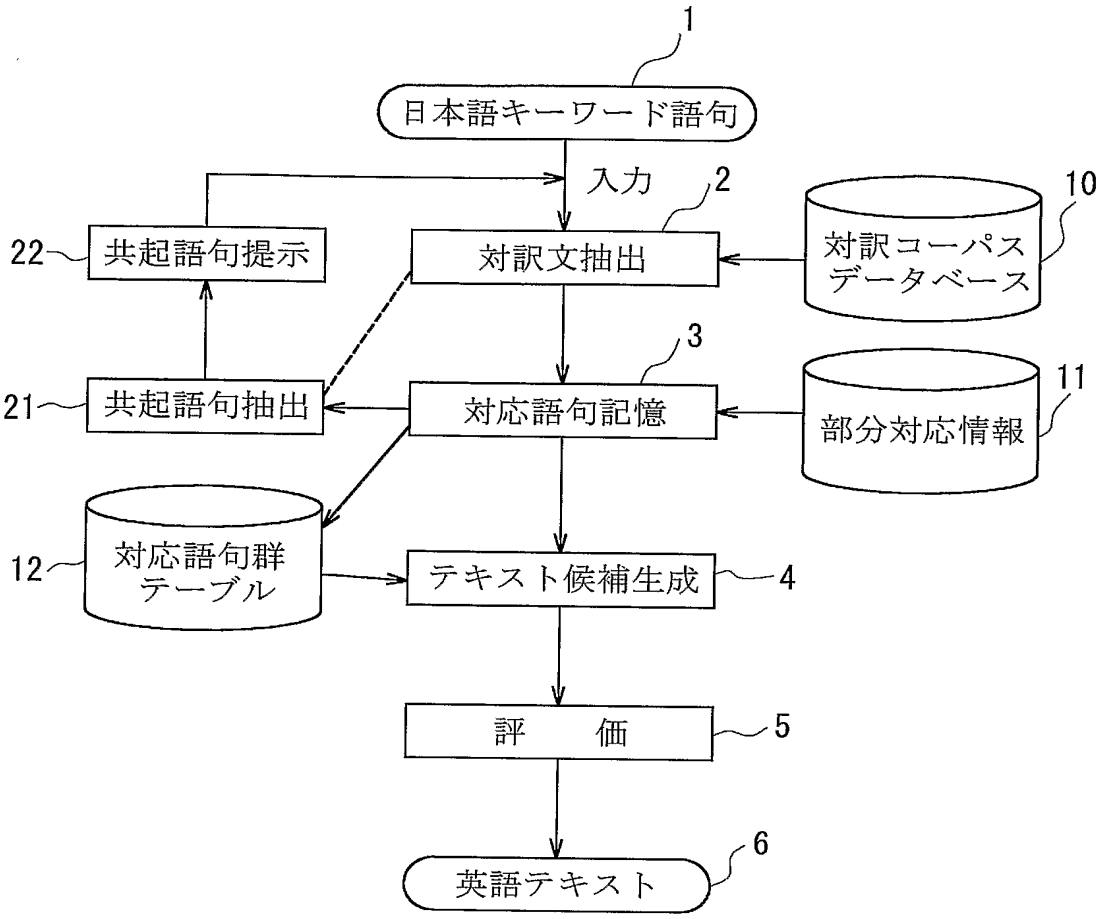
第２図


1
日本語キーワード語句
入力
2
対訳文抽出
10
対訳コーパス
データベース
12
対応語句群
テーブル
3
対応語句記憶
11
部分対応情報
4
テキスト候補生成
5
評　価
6
英語テキスト

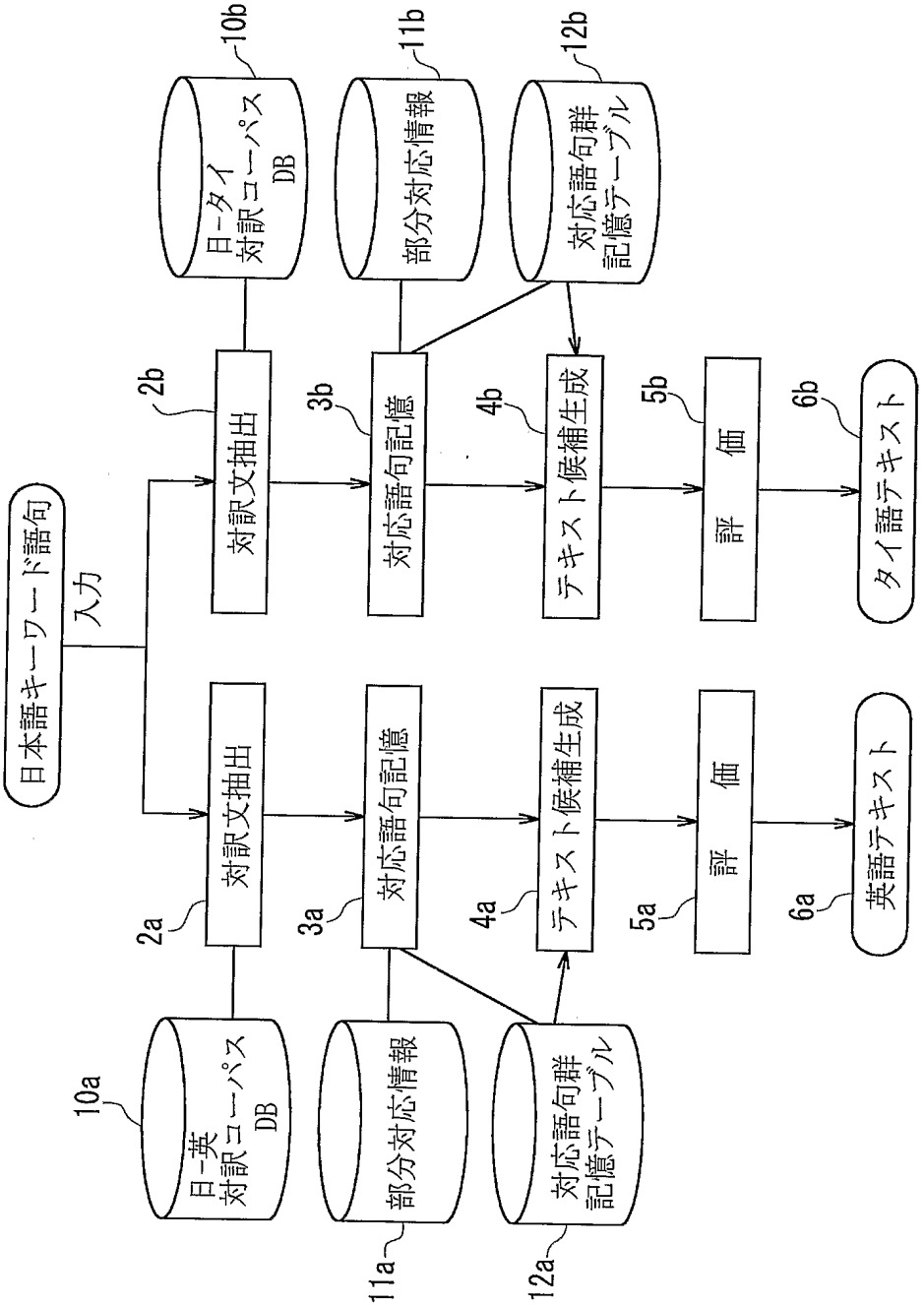
第３図


1
日本語キーワード語句
入力
2
対訳文抽出
対訳コーパス
データベース
10
20
日本語キーワード
絞り込み
対応語句群
テーブル
対応語句記憶
3
部分対応
情報
12
11
テキスト候補生成
4
評　価
5
英語テキスト
6

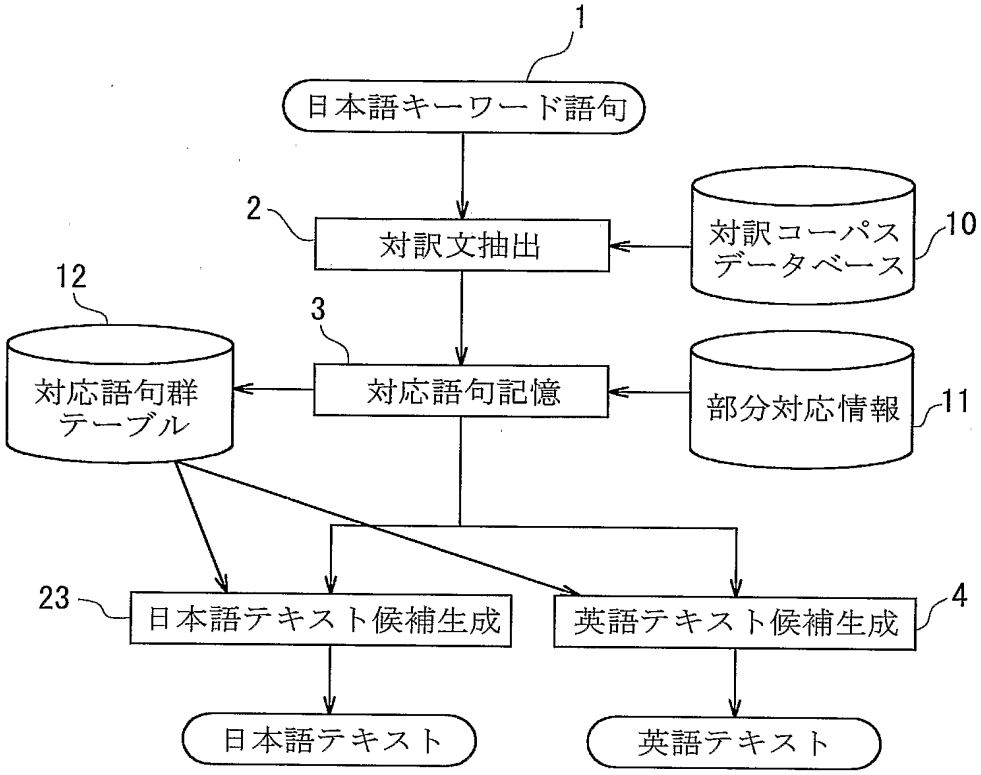
第４図


1
日本語キーワード語句
入力
2
10
対訳コーパス
データベース
22
共起語句提示
対訳文抽出
3
11
21
共起語句抽出
対応語句記憶
部分対応情報
12
対応語句群
テーブル
テキスト候補生成
4
評　価
5
6
英語テキスト

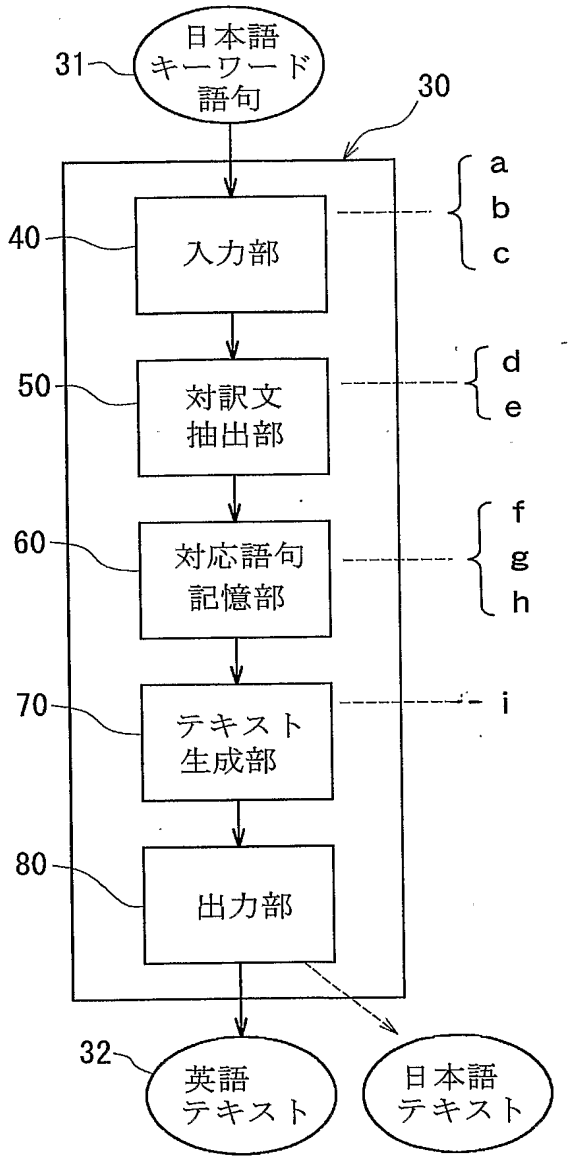
第５図


日本語キーワード語句
入力
対訳文抽出
2a
対応語句記憶
3a
テキスト候補生成
4a
評　価
5a
英語テキスト
6a
日-英
対訳コーパス
DB
10a
部分対応情報
11a
対応語句群
記憶テーブル
12a
対訳文抽出
2b
対応語句記憶
3b
テキスト候補生成
4b
評　価
5b
タイ語テキスト
6b
日-タイ
対訳コーパス
DB
10b
部分対応情報
11b
対応語句群
記憶テーブル
12b

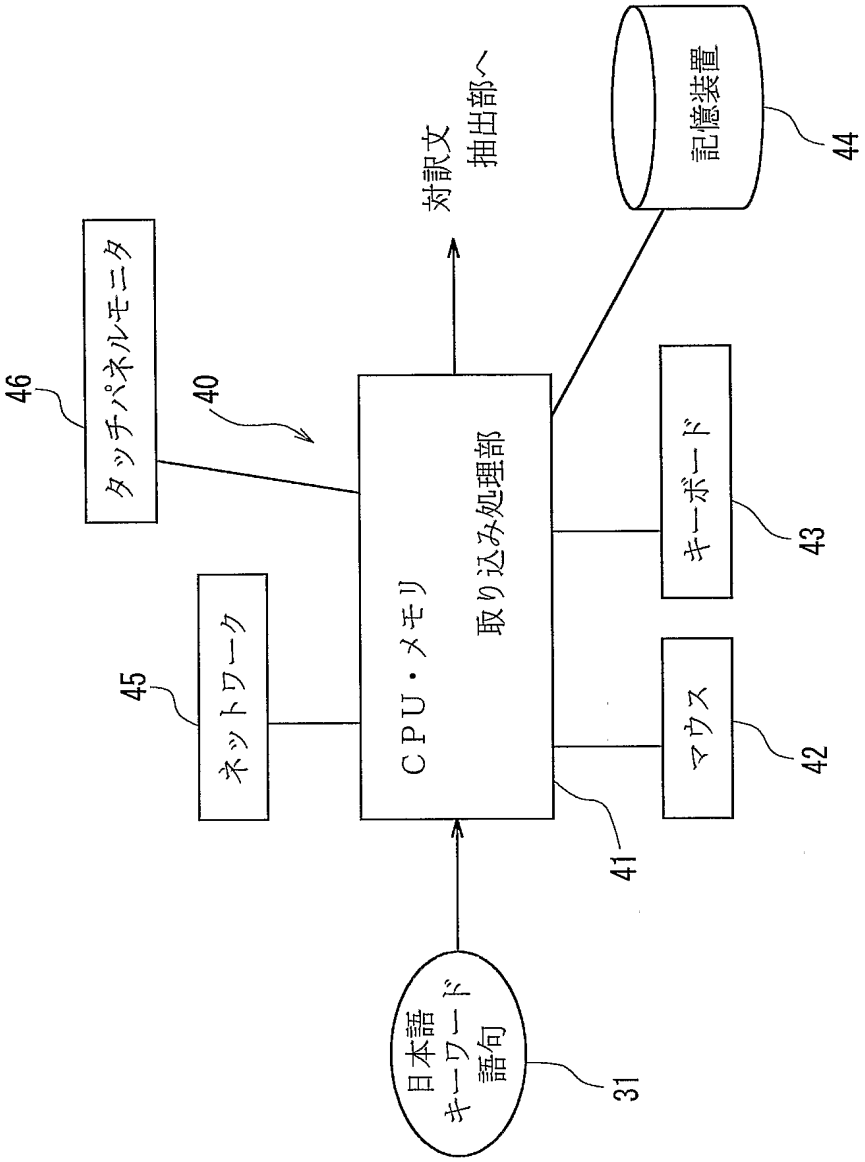
第６図


1
日本語キーワード語句
2
対訳文抽出
対訳コーパス
データベース
10
12
3
対応語句群
テーブル
対応語句記憶
部分対応情報
11
23
日本語テキスト候補生成
英語テキスト候補生成
4
日本語テキスト
英語テキスト

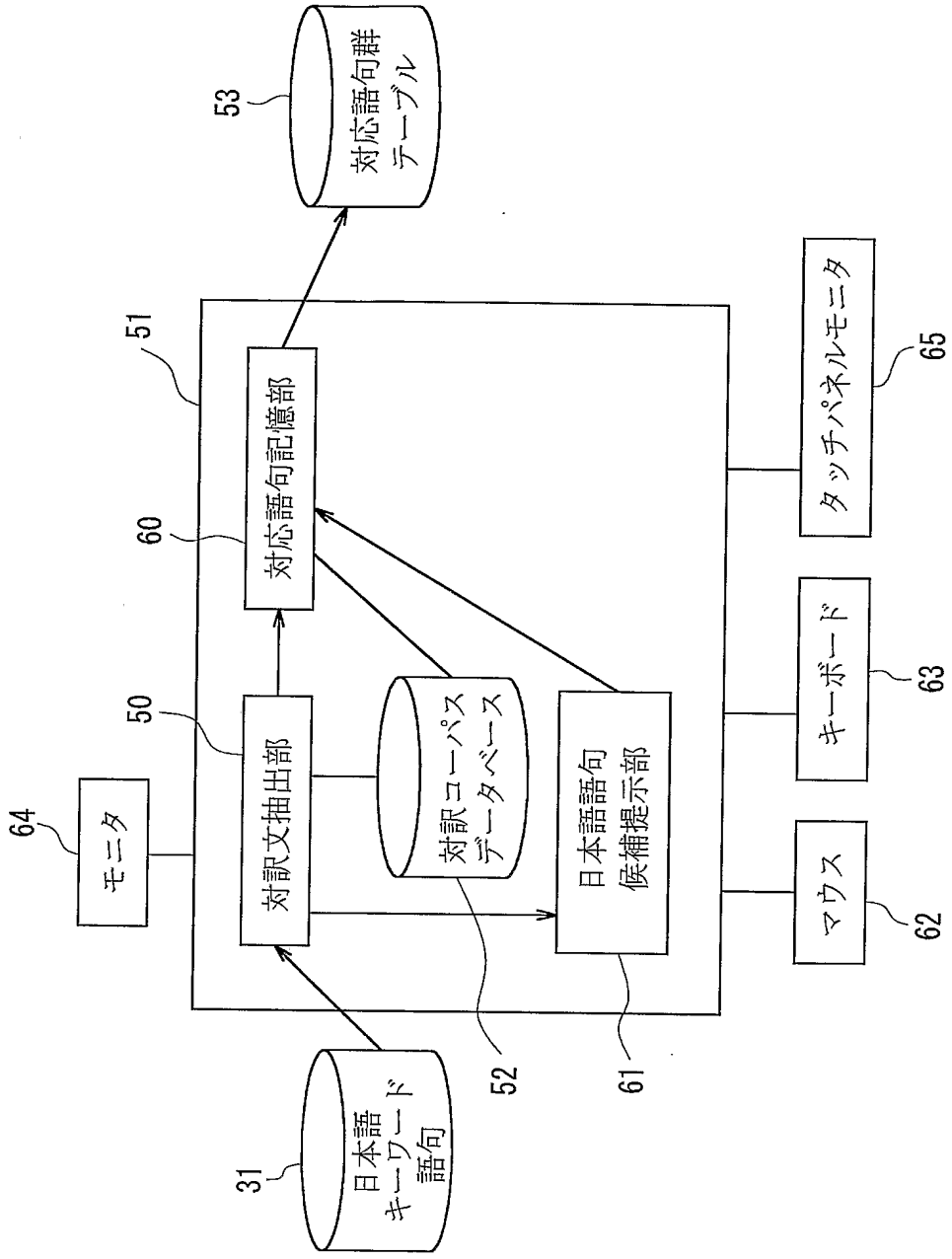
第７図


日本語
キーワード
語句
31
30
a
b
c
40
入力部
d
e
50
対訳文
抽出部
f
g
h
60
対応語句
記憶部
i
70
テキスト
生成部
80
出力部
32
英語
テキスト
日本語
テキスト

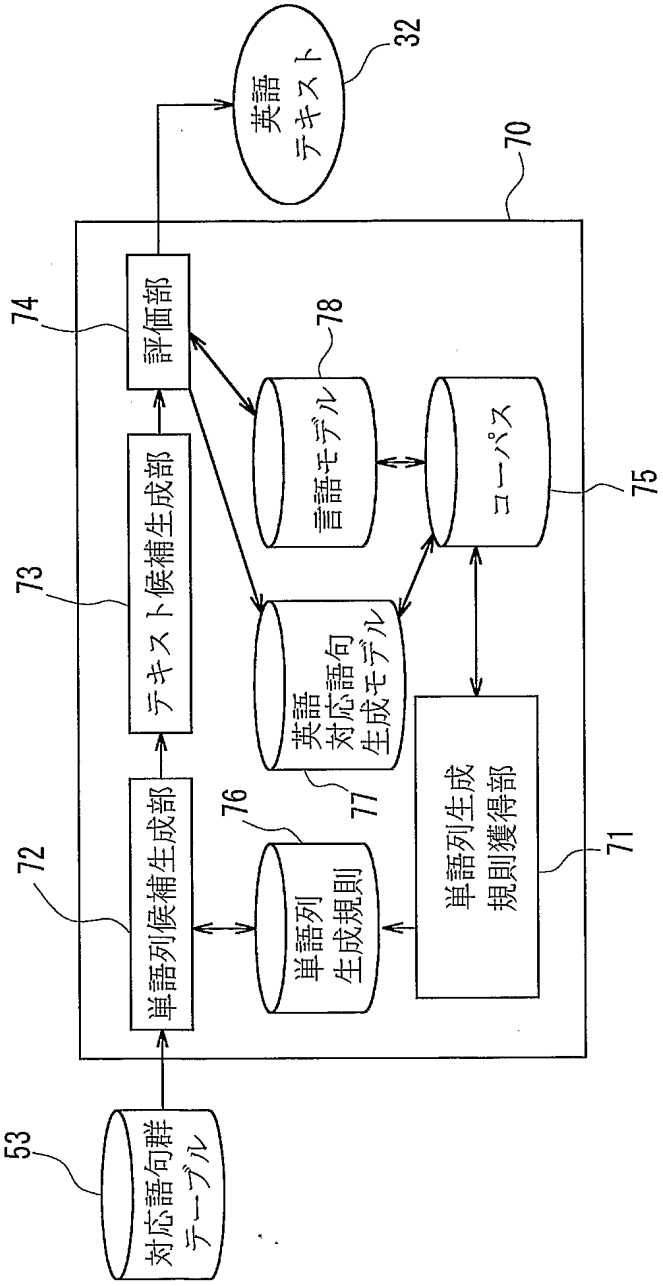
第８図


対訳文
抽出部へ
記憶装置
44
タッチパネルモニタ
46
40
CPU・メモリ
取り込み処理部
キーボード
43
ネットワーク
45
マウス
42
41
日本語
キーワード
語句
31

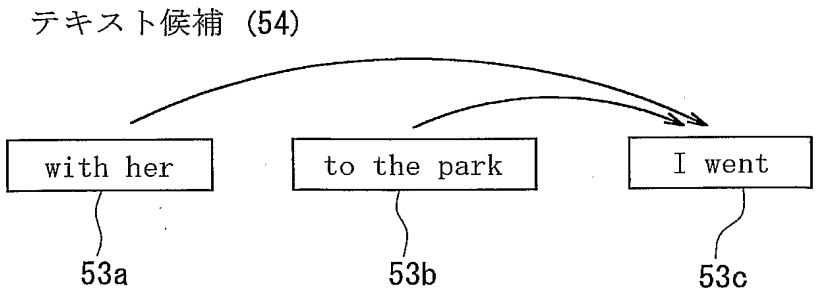
第９図


53
対応語句群テーブル
51
対応語句記憶部
60
50
対訳文抽出部
対訳コーパスデータベース
日本語語句候補提示部
タッチパネルモニタ
65
キーボード
63
マウス
62
モニタ
64
日本語キーワード語句
31
52
61

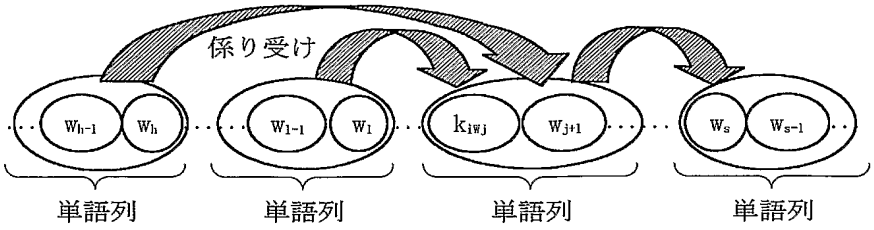
第１０図


英語テキスト
32
70
評価部
74
78
言語モデル
コーパス
75
テキスト候補生成部
73
英語対応語句生成モデル
77
単語列候補生成部
72
76
単語列生成規則
単語列生成規則獲得部
71
対応語句群テーブル
53

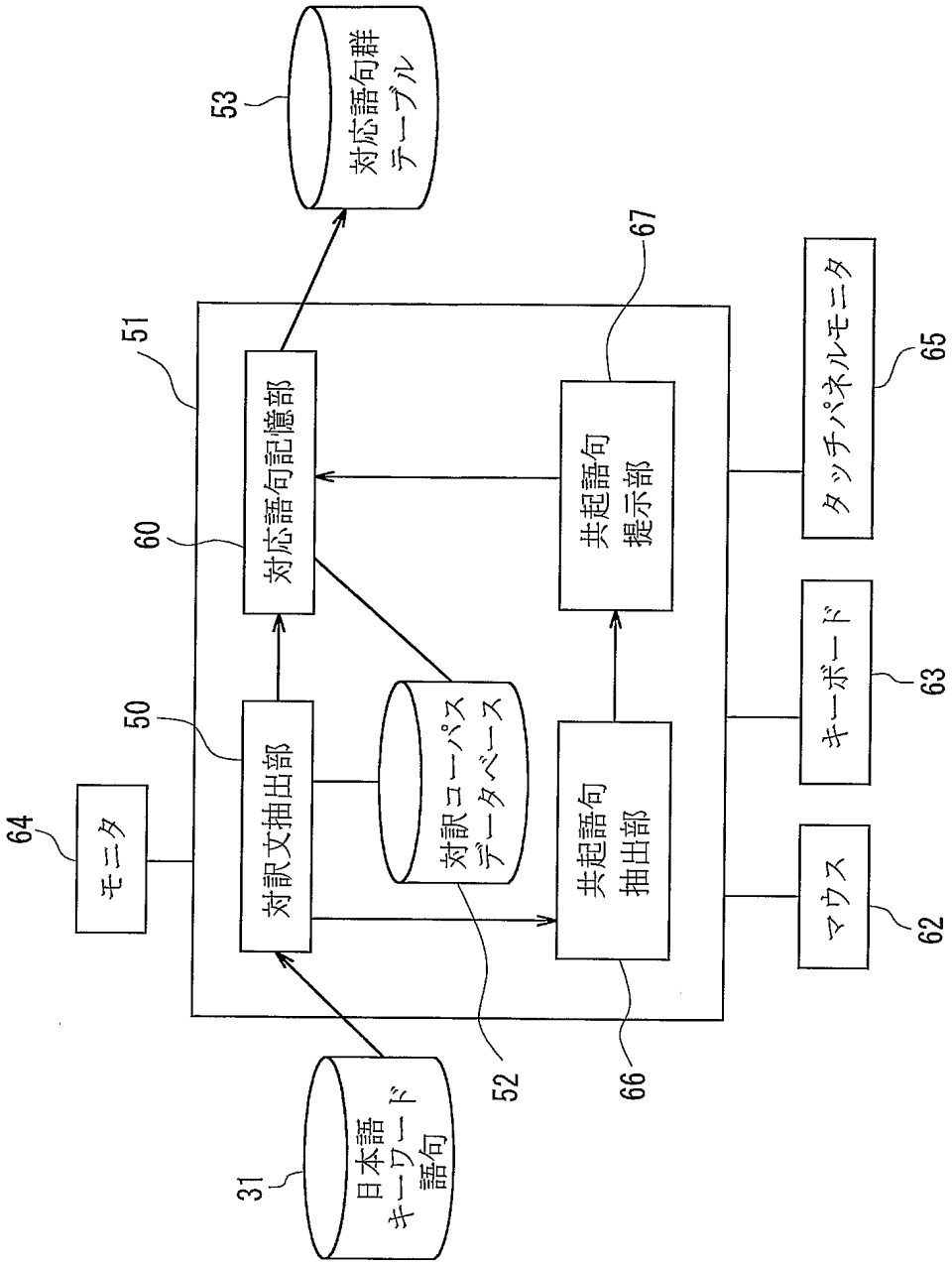
第１１図



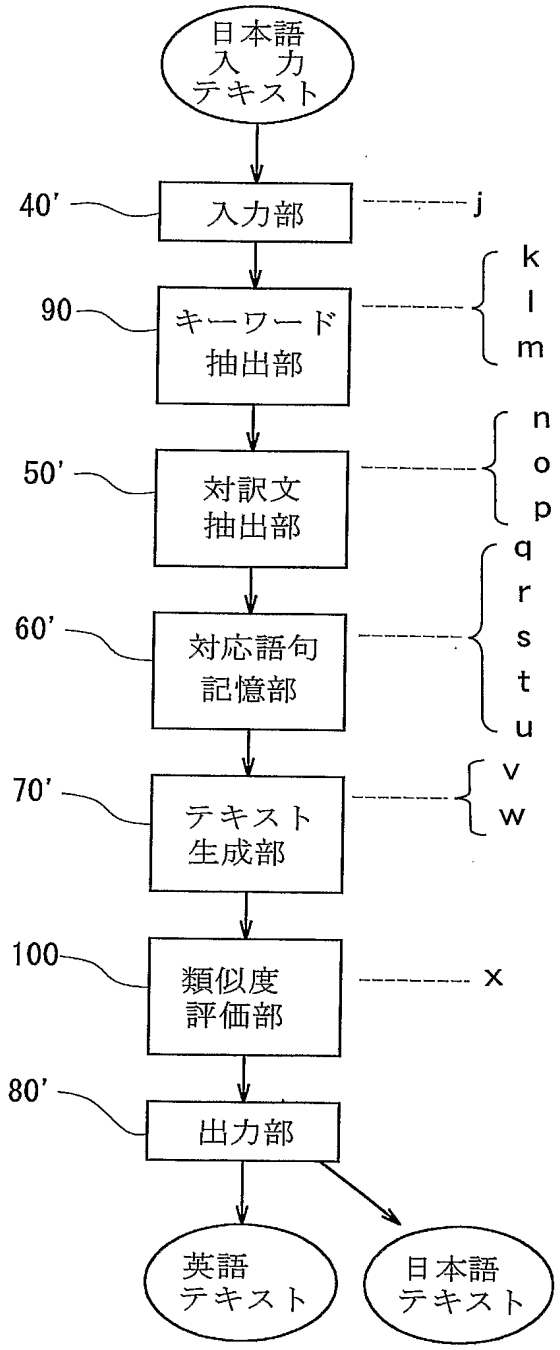
第１２図



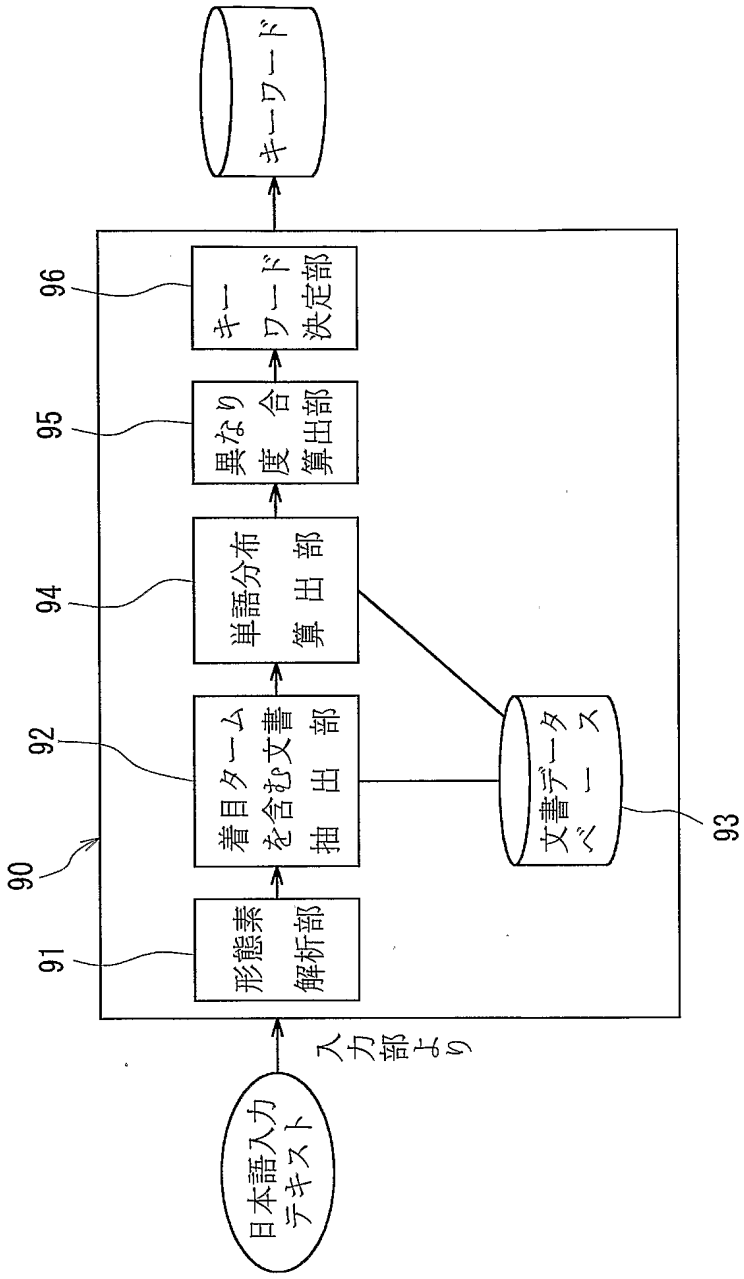
第１３図


対応語句群テーブル
53
対応語句記憶部
51
60
対訳文抽出部
50
モニタ
64
対訳コーパスデータベース
52
共起語句抽出部
66
共起語句提示部
67
日本語キーワード語句
31
タッチパネルモニタ
65
キーボード
63
マウス
62

第１４図


日本語
入　力
テキスト
40'
入力部
j
90
キーワード
抽出部
k
l
m
50'
対訳文
抽出部
n
o
p
60'
対応語句
記憶部
q
r
s
t
u
70'
テキスト
生成部
v
w
100
類似度
評価部
x
80'
出力部
英語
テキスト
日本語
テキスト

第１５図


キーワード
96
キーワード決定部
95
異なり度合算出部
94
単語分布算出部
92
着目タームを含む文書抽出部
91
形態素解析部
90
93
文書データベース
入力部より
日本語入力テキスト

符号の説明

３０　テキスト生成装置
３１　日本語キーワード語句
３２　英語テキスト
４０　入力部
５０　対訳文抽出部
６０　対応語句記憶部
７０　テキスト生成部
８０　出力部
ａ，ｋ　「彼女」
ｂ，ｌ　「公園」
ｃ，ｍ　「行く」
ｄ，ｎ　「公園へ行った。／ I went to the park. 」
ｅ，ｏ　「彼女と百貨店へ行った。／ I went to the department store with her. 」
ｆ，ｑ　公園へ／ to the park
ｇ，ｒ　行った／ I ・・・went
ｈ，ｓ　彼女と／ with her
ｉ　「彼女と公園へ行った。／ I went to the park with her. 」
ｊ　「彼女は公園へ行った。」
ｐ　「彼女は動物園へ行った。／ She went to the zoo. 」
ｔ　彼女は・・・行った／ She went ・・・
ｕ　動物園へ／ to the zoo
ｖ　①「彼女と公園へ行った。／ I went to the park with her. 」
ｗ　②「彼女は動物園へ行った。／ She went to the zoo. 」
ｘ　②の類似度が高い

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2005/001636

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl[7] G06F17/28

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl[7] G06F17/21-30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1922-1996 Toroku Jitsuyo Shinan Koho 1994-2005
Kokai Jitsuyo Shinan Koho 1971-2005 Jitsuyo Shinan Toroku Koho 1996-2005

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JSTPlus FILE(JOIS), WPI, INSPEC(DIALOG)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2003-196280 A (Communications Research Laboratory), 11 July, 2003 (11.07.03), Claims & WO 2003/056451 A1	1-14
A	JP 2003-271592 A (Communications Research Laboratory), 26 September, 2003 (26.09.03), Claims & WO 2003/079224 A1	1-14
A	JP 05-250407 A (Hitachi, Ltd.), 28 September, 1993 (28.09.93), Par. No. [0011] (Family: none)	1-14

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
21 April, 2005 (21.04.05)	17 May, 2005 (17.05.05)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

国際調査報告

国際出願番号　PCT／JP2005／001636

A.　発明の属する分野の分類（国際特許分類（IPC））

Int. Cl⁷ G06F17／28

B.　調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int. Cl⁷ G06F17／21－30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922－1996年
日本国公開実用新案公報	1971－2005年
日本国登録実用新案公報	1994－2005年
日本国実用新案登録公報	1996－2005年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

JSTPlusファイル（JOIS），WPI，INSPEC（DIALOG）

C.　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
A	JP 2003-196280 A(独立行政法人通信総合研究所)2003.07.11，特許請求の範囲　&　WO 2003/056451 A1	1－14
A	JP 2003-271592 A(独立行政法人通信総合研究所)2003.09.26，特許請求の範囲　&　WO 2003/079224 A1	1－14
A	JP 05-250407 A(株式会社日立製作所)1993.09.28，第11段落（ファミリーなし）	1－14

☐ C欄の続きにも文献が列挙されている。

☐ パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの

「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）

「O」口頭による開示、使用、展示等に言及する文献

「P」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献

「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「&」同一パテントファミリー文献

国際調査を完了した日 21.04.2005	国際調査報告の発送日　17.05.2005
国際調査機関の名称及びあて先 日本国特許庁（ISA／JP） 郵便番号100－8915 東京都千代田区霞が関三丁目4番3号	特許庁審査官（権限のある職員）　5M　9459 和田　財太 電話番号　03－3581－1101　内線　3597

様式PCT／ISA／210（第2ページ）（2004年1月）